Amendments to the Claims:

1. (previously presented): A document descriptor determination method comprising the steps of:

generalizing input sequences associated with a document to develop general sequences, said input sequences reflecting the structure of a document;

factoring said input sequences and said general sequences to develop factored sequences;

selecting a document descriptor from said input sequences, said general sequences, and said factored sequences using minimum descriptor length (MDL) principles.

2. (original): The method of claim 1, wherein said selecting step comprises the steps of:

encoding said input sequences, said general sequences, and said factored sequences; and

selecting a document descriptor which encompasses all of said input sequences and exhibits a minimum MDL cost.

3. (original): The method of claim 2, wherein said encoding step employs an algorithm which applies a set of rules comprising:

$seq(D,s) = \varepsilon$ if D=s, if D does not contain metacharacters;

$seq(D_1...D_k, s_1...s_k) = seq(D_1,s_1)...seq(D_k,s_k)$;

$seq(D_1|...|D_m,s) = i\ seq(D_i,s)$;

$seq(D^*,s_1...s_k) = \{k\ seq(D,s_1)...seq(D,s_k)$ if $k>0$; 0 otherwise$\}$;

wherein D is a sequence of symbols, s is a sequence, and i is an index of a regular expression that the corresponding sequence s matches, wherein log m bits are needed to encode index i.

4. (original): The method of claim 3, wherein said minimum MDL cost is determined by employing an algorithm to solve a facility location problem (FLP), said FLP modified to compute said minimum MDL cost of potential document descriptors.

5. (original): The method of claim 4, wherein said document descriptor is a document type descriptor (DTD), and said document is an eXtensible Markup Language (XML) document.

6. (original): The method of claim 5, wherein said minimum MDL cost comprises summing a first length of bits describing the DTD and a second length of bits for encoding the sequences.

7. (previously presented): A document descriptor determination method comprising the steps of:

generalizing input sequences to develop general sequences, said input sequences reflecting the structure of data within a document;

selecting a document descriptor from said input sequences and said general sequences using minimum descriptor length (MDL) principles.

8. (original): The method of claim 7, wherein said selecting step comprises the steps of:

encoding said input sequences and said general sequences; and

selecting a document descriptor which encompasses all of said input sequences and exhibits a minimum MDL cost.

9. (original): The method of claim 8, wherein said encoding step employs an algorithms which applies a set of rules comprising:

$seq(D,s) = \varepsilon$ if D=s, if D does not contain metacharacters;

$seq(D_1...D_k, s_1...s_k) = seq(D_1,s_1)...seq(D_k,s_k)$, if D is a concatenation of $D_1...D_k$;

$seq(D_1|...|D_m,s) = i\ seq(D_i,s)$;

$seq(D^*,s_1...s_k) = \{k\ seq(D,s_1)...seq(D,s_k)$ if $k>0$; 0 otherwise$\}$;

wherein D is a sequence of symbols, s is a sequence, and i is an index of a regular expression that the corresponding sequence s matches, wherein log m bits are needed to encode index i.

10. (original): The method of claim 9, wherein said minimum MDL cost is determined by employing an algorithm to solve a facility location problem (FLP), wherein said FLP is modified to compute said minimum MDL cost of potential document descriptors.

11. (original): The method of claim 10, wherein said document descriptor is a document type descriptor (DTD), and said document is an eXtensible Markup Language (XML) document.

12. (original): The method of claim 11, wherein said minimum MDL cost comprises summing a first length of bits describing the DTD and a second length of bits for encoding the sequences.

13. (previously presented): The method of claim 7, further comprising the step of:

factoring said input sequences and said general sequences to develop factored sequences, wherein said factored sequences are available for said step of selecting.

14-17 (cancelled)

18. (previously presented): A document descriptor determination method comprising the steps of:

generalizing input sequences, said generalizing step comprising the steps of:

discovering OR patterns among said input sequences, and

discovering sequence patterns among said input sequences and OR patterns; and

selecting a document descriptor from said input sequences and said general sequences.

19. (original): The method of claim 18, wherein said discovering OR patterns step comprises the step of partitioning said input sequences.

20. (original): The method of claim 19, further comprising the steps of:

factoring said input sequences and said general sequences to develop factored sequences, wherein said factored sequences are available to said step of selecting.

21. (original): The method of claim 20, wherein said step of selecting employs minimum descriptor length (MDL) principles.

22. (original): The method of claim 21, wherein said document descriptor is a document type descriptor (DTD) and said document is an eXtensible Markup Language (XML) document.